



# Python - All A Scientist Needs

[openwetware.org/wiki/python](http://openwetware.org/wiki/python)

**Julius B. Lucks**

**Miller Institute and Berkeley Center for Synthetic Biology**

**University of California, Berkeley**



# Python



- ❖ Python is a complete scientific programming platform
- ❖ Python promotes good scientific practice

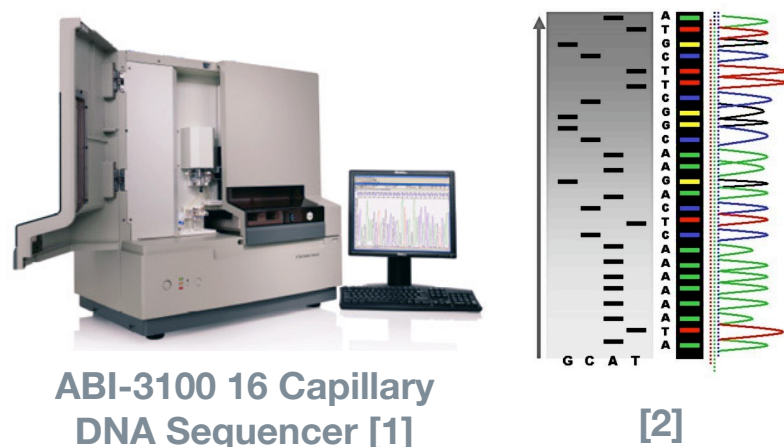


# The Scientist's Dilemma



## Hypothesis

### Raw Data

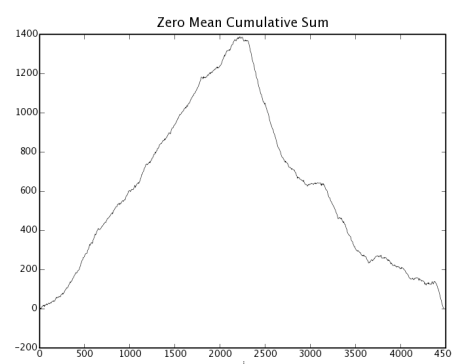


### Data Processing

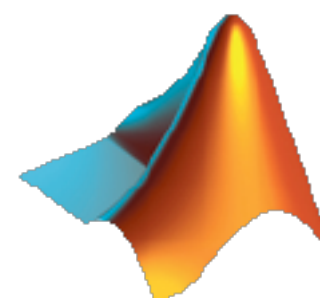
ATGC...

### Data Management

ATGC...  
ATGC...  
ATGC...  
ATGC...  
ATGC...  
ATGC...  
ATGC...  
ATGC...



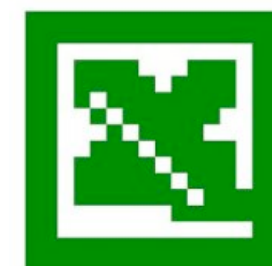
### Data Presentation



Matlab



Mathematica



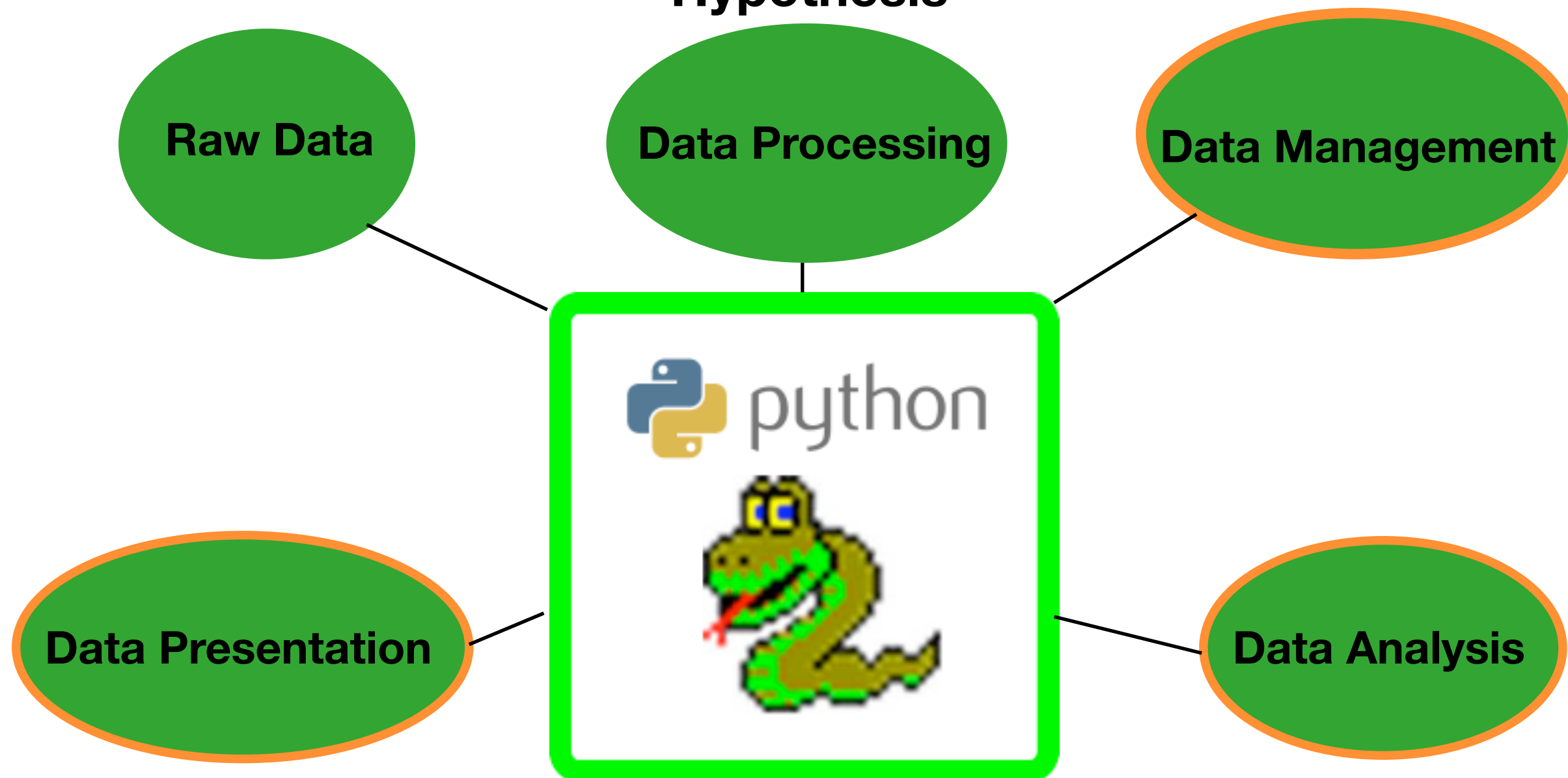
Excel

### Data Analysis

# The Scientist's Dilemma



## Hypothesis

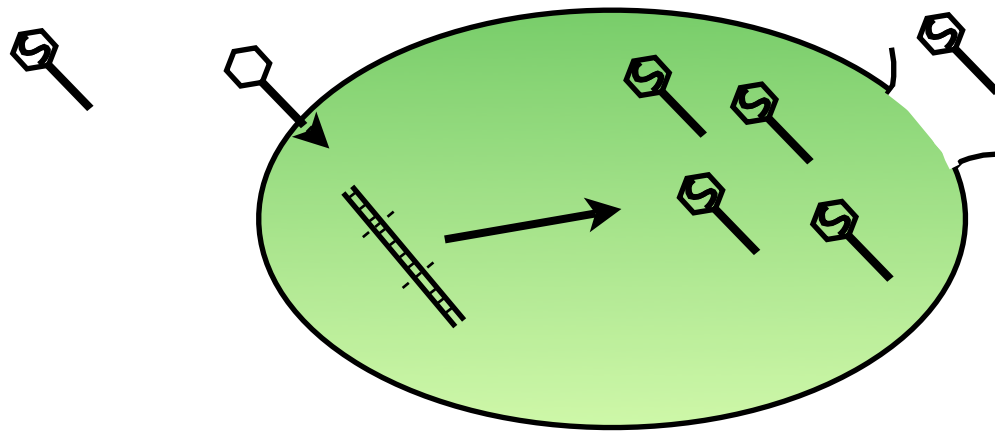


# Case Study



## Comparative Genomics of Viruses

### Viral Life Cycle



Protein ..... Ser Pro Thr Ala .....

mRNA1 ..... UCUCCUACUGCU .....

mRNA2 ..... UCCCCCACCGCC .....

⋮

### The Genetic Code

UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp
CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }
AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }
GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }

[3]

**Synonymous spellings** give us a way to detect selection pressures via **preferred spellings**

# Comparative Genomics Tasks



- ❏ Download, store, and parse genome files
- ❏ Visualize sequences with 'genome landscape' plots
- ❏ Draw random genomes to compare to the sequenced genome





# GenBank



http://www.ncbi.nlm.nih.gov/sites/entrez?term=Bacteriophage&cmd=Search&db=nucleotide&QueryKey=10

NCBI

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search CoreNucleotide for Bacteriophage Go Clear Save Search

Limits Preview/Index History Clipboard Details

Found 11710 nucleotide sequences. CoreNucleotide [10342] EST [992] GSS [376]

Display Summary Show 20 Sort by Send to

All: 10342 Bacteria: 3127 RefSeq: 1959 mRNA: 441

Items 1 - 20 of 10342 Page 1 of 518 Next

1: CP000958 Reports Links  
Burkholderia cenocepacia MC0-3 chromosome 1, complete sequence  
gil169814598|gb|CP000958.1|[169814598]

2: NC\_010410 Reports Links  
Acinetobacter baumannii AYE, complete genome  
gil169794206|ref|NC\_010410.1|[169794206]

3: NC\_010404 Reports Links  
Acinetobacter baumannii plasmid p3ABAYE, complete sequence  
gil169786889|ref|NC\_010404.1|[169786889]

4: NC\_009348 Reports Links  
Aeromonas salmonicida subsp. salmonicida A449, complete genome  
gil145297124|ref|NC\_009348.1|[145297124]

5: AE005174 Reports Links  
Escherichia coli O157:H7 EDL933, complete genome  
gil56384585|gb|AE005174.2|[56384585]

6: AE014073 Reports Links  
Shigella flexneri 2a str. 2457T, complete genome  
gil30043918|gb|AE014073.1|[30043918]

Top Organisms [Tree]  
unidentified (1604)  
synthetic construct (818)  
Escherichia coli (440)  
Candidatus Hamiltonella defensa (228)  
Enterobacteria phage T7 (173)  
All other taxa (6969)  
More...



# Biopython



## Task: Download and Parse GenBank files

```

LOCUS      NC_001416                48502 bp    DNA        linear    PHG 28-NOV-2007
DEFINITION Enterobacteria phage lambda, complete genome.
ACCESSION  NC_001416
VERSION    NC_001416.1  GI:9626243
...
FEATURES             Location/Qualifiers
     source            1..48502
                        /organism="Enterobacteria phage lambda"
                        /specific_host="Escherichia coli"
                        /db_xref="taxon:10710"
     gene              191..736
                        /gene="nu1"
                        /db_xref="GeneID:2703523"
     CDS               191..736
                        /gene="nu1"
                        /codon_start=1
                        /transl_table=11
                        /product="DNA packaging protein"
                        /protein_id="NP_040580.1"
                        /db_xref="GI:9626244"
                        /db_xref="GeneID:2703523"
                        /translation="MEVNKKQLADIFGASIRTIONWQEQGMPVLRGGGKGNEVLYDSA
AVIKWYAERDAEIE NEKLRREVEELRQASEADLQPGTIEYERHRLTRAQADAQELKNA
...
ORIGIN
      1  gggcggcgac ctcgcgggtt ttcgctat tt atgaaaattt tccggtttta ggcgtttccg
     61  ttctttcttcg tcataactta atgtttttat ttaaaatacc ctctgaaaag aaaggaaacg
  
```

Record Information

Genomic Features

DNA Sequence





# Biopython



## Task: Download and Parse GenBank files

```
from Bio import GenBank
```

Biopython modules for GenBank files

```
def download(accession_list):
```

Download multiple GenBank files

```
    try:
```

```
        handle = GenBank.download_many(accession_list)
```

```
    except:
```

```
        ...
```

```
    genbank_strings = handle.read().split('//\n')
```

Split and save

```
    for i in range(len(genbank_strings)):
```

```
        #Save record in file
```

```
        ...
```

Usage

```
import genbank
```

```
genbank.download(['NC_001416'])
```



# Matplotlib



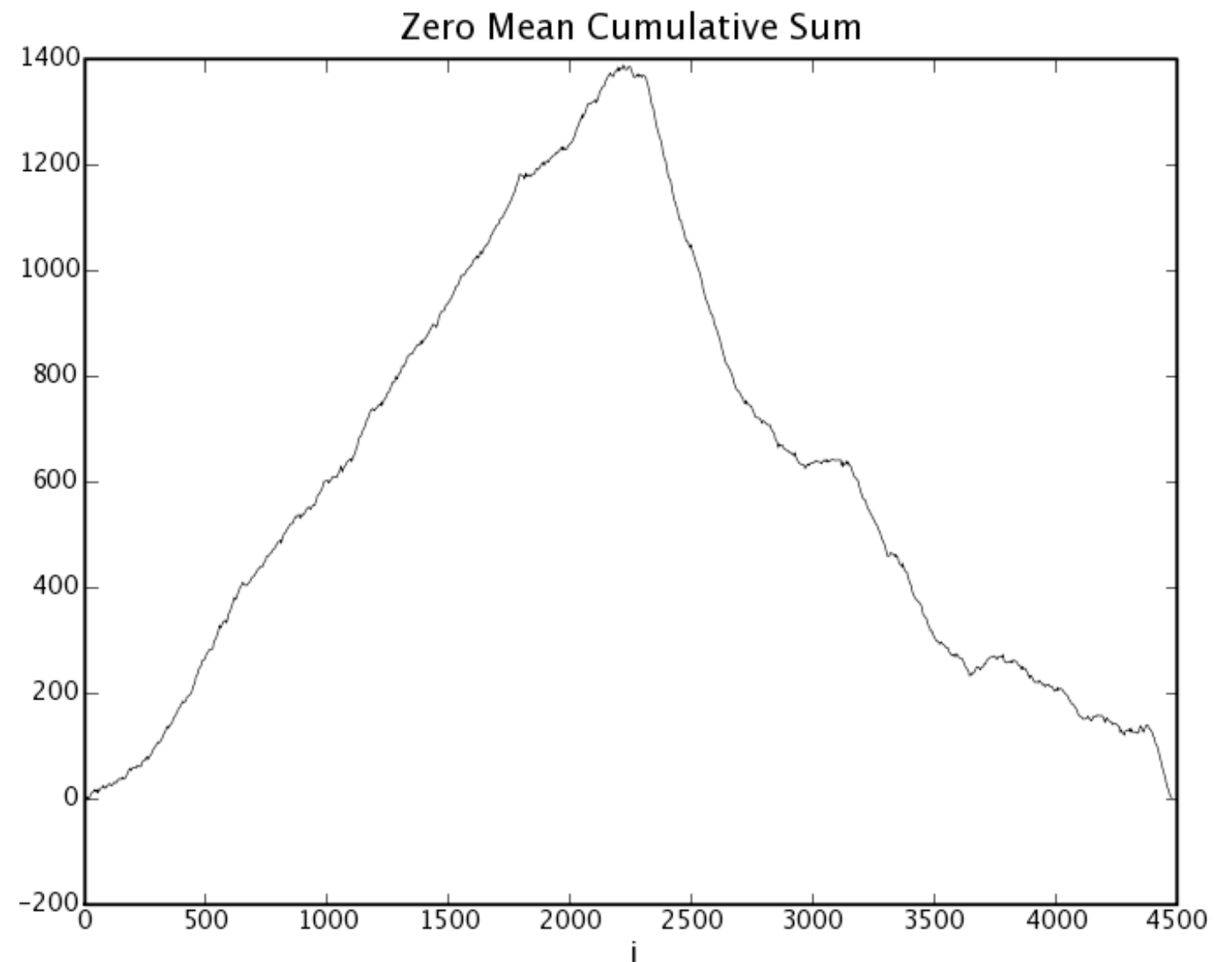
## Task: Plot Genome Landscapes

DNA .....TCTAAGACTGCT.....

$\eta(i)$  .....010001010110.....

$$\eta(i) = 1(\text{GC}), 0(\text{AT})$$

$$F(m) = \sum_{i=0}^m (\eta(i) - \bar{\eta})$$



# Matplotlib



## Task: Plot Genome Landscapes

```
import fileinput
import numpy
from matplotlib import pylab
```

Standard Library  
Numerical  
Matlab-style plotting



```
def plot(filename):
```

```
    numbers = []
    for line in fileinput.input(filename):
        numbers.append(float(line.split('\n')[0]))
```

Parse input file



```
    mean = numpy.mean(numbers)
    cumulative_sum = numpy.cumsum([number - mean for number in numbers])
```

Cumulative sum



```
    pylab.plot(cumulative_sum[0::10], 'k-')
```

Plot



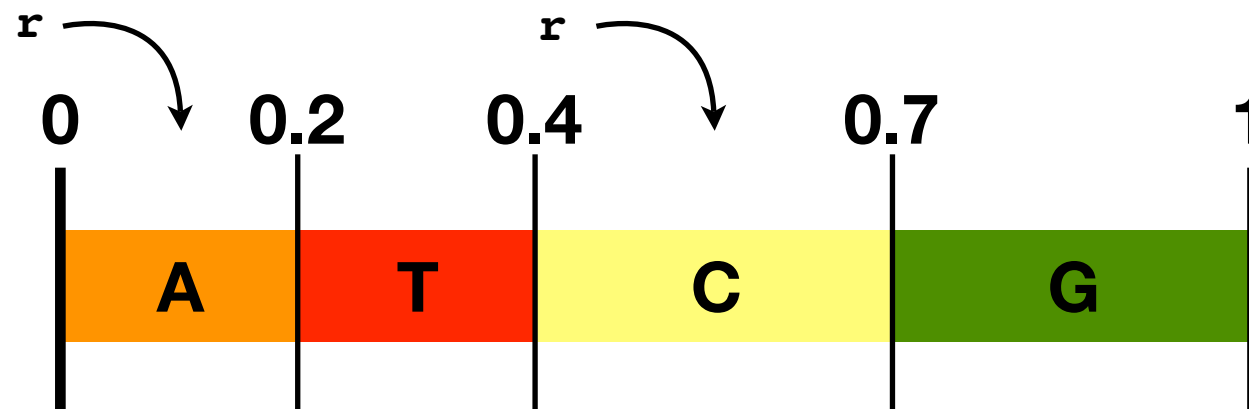
```
...
```



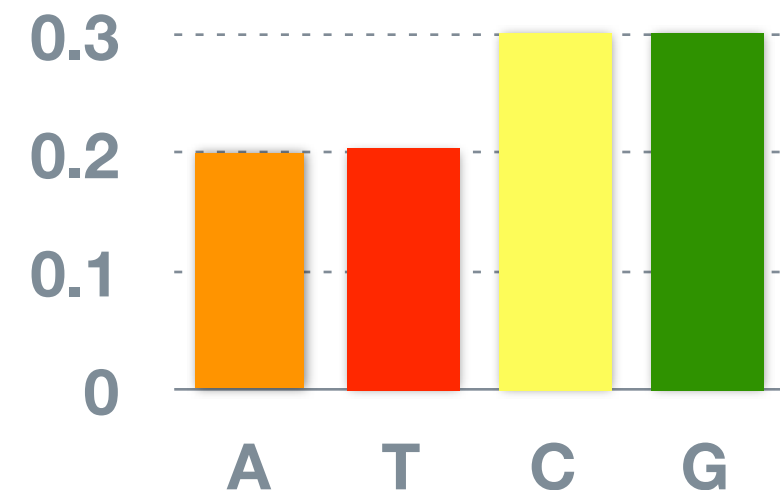
# SWIG



## Task: Speed up the code!



Draw random genomes according to distribution of nucleotide frequencies



```
import random
def seed():
    random.seed()
def draw(distribution):
    sum = 0
    r = random.random()
    for i in range(0, len(distribution)):
        sum += distribution[i]
        if r < sum:
            return i
```

← Standard library module

← Seed random number generator

← Draw according to input distribution

# SWIG



## Step 1: Create a C implementation

```
#include <stdlib.h>
#include <stdio.h>
#include <time.h>

void seed() {
    srand((unsigned) time(NULL) * getpid());
}

int draw(float distribution[4]) {
    float r= ((float) rand() / (float) RAND_MAX);
    float sum = 0.;
    int i = 0;
    for(i = 0; i < 4; i++) {
        sum += distribution[i];
        if (r < sum) {
            return i;
        }
    }
}
```



# SWIG



## 🔧 Step 2: Create a SWIG Interface File

```
%module c_discrete_distribution
%typemap(in) float[4](float temp[4]) {
    int i;
    if (PyList_Check($input)) {
        PyObject* input_to_tuple = PyList_AsTuple($input);
        if (!PyArg_ParseTuple(input_to_tuple, "ffff", temp, temp+1, temp+2, temp+3)) {
            PyErr_SetString(PyExc_TypeError, "tuple must have 4 elements");
            return NULL;
        }
        $1 = &temp[0];
    } else {
        PyErr_SetString(PyExc_TypeError, "expected a tuple.");
        return NULL;
    }
}

void seed();
int draw(float distribution[4]);
```

Module Name

Convert Python List to C-Array

Declare interface



# SWIG



```
import discrete_distribution
```

```
...
```

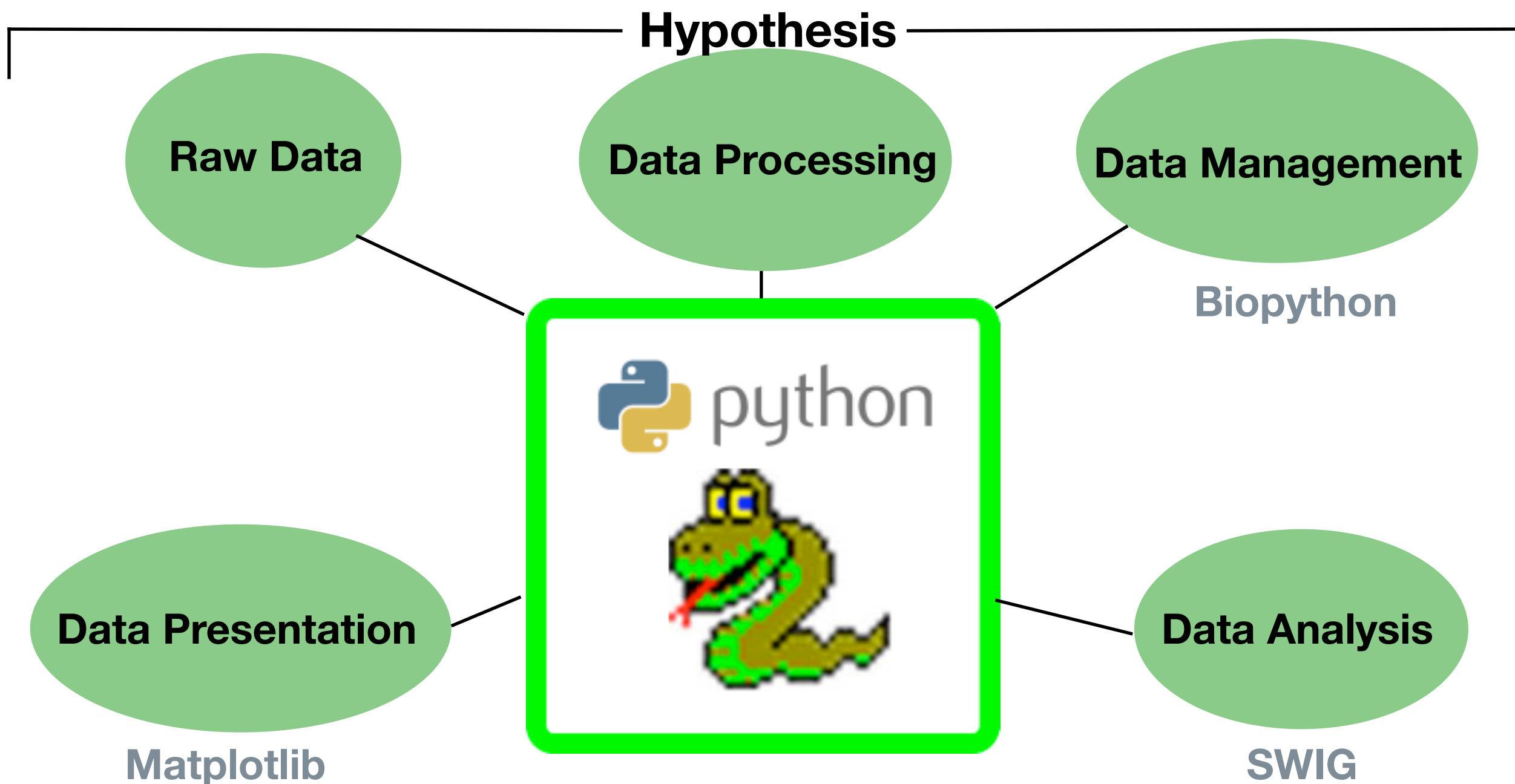


```
import c_discrete_distribution as discrete_distribution
```

```
...
```



# Conclusions



# References/Credits



Thank you to the Miller Institute for Basic Scientific Research

openwetware.org/wiki/python

- [1] [http://ccg.biology.uiowa.edu/equipment\\_ABI3100.php](http://ccg.biology.uiowa.edu/equipment_ABI3100.php)
- [2] (Abizar Lakdawalla)  
[http://en.wikipedia.org/wiki/Image:Radioactive\\_Fluorescent\\_Seq.jpg](http://en.wikipedia.org/wiki/Image:Radioactive_Fluorescent_Seq.jpg)
- [3] [http://www.mun.ca/biology/scarr/MGA2\\_03-20.html](http://www.mun.ca/biology/scarr/MGA2_03-20.html)
- [4] <http://www.ncbi.nlm.nih.gov/Genbank/>

